引用格式：朱明婷, 徐崇利. 人工智能伦理的国际软法之治：现状、挑战与对策. 中国科学院院刊, 2023, 38(7): 1037-1049

Zhu M T, Xu C L. International soft law governance of artificial intelligence ethics: Current situation, challenges and countermeasures. Bulletin of Chinese Academy of Sciences, 2023, 38(7): 1037-1049

人工智能伦理的国际软法之治：现状、挑战与对策

朱明婷* 徐崇利

厦门大学 法学院 厦门 361005

摘要 人工智能技术不仅能快速赋能经济社会发展，也可能引发诸多与人工智能技术本身特征和发展高度相关的伦理问题。国际软法因具有灵活高效、适用成本低，能填补硬法空白，以及方便区分治理、分层应对伦理问题的优势，其在人工智能伦理治理领域的勃兴几乎是必然的。在该领域国际软法发达、硬法落后的现状下，面对国际软法主体间合作不稳定、有时得不到有效实施的治理挑战，治理模式逐渐向软硬兼备、软法“硬化”转变，以提高软法约束力与执行可能性。建议构造国际软硬法混合治理的“中心—外围”模式、构建间接执行机制，以完善人工智能伦理的国际软法治理对策。

关键词 人工智能伦理，国际软法，混合治理，间接执行机制

CSTR 32128.14.CASbulletin.20221003002

正如过去工业时代的蒸汽机和电力一样，随着数据的积累、算法的进步，以及算力的提高，人工智能（AI）技术正重塑人类世界。ChatGPT（生成式预训转换机器人）的诞生就是智能机器人发展的一个阶段性进步。但是，在取代传统搜索引擎成为人类工作和生活更为便利的“百科全书”的同时，ChatGPT也使得人们产生诸如学生利用AI作弊、数据泄露、后台技术操控者价值渗透等伦理隐忧。

从理论上讲，人工智能可能存在技术失控和技术的非正当应用风险。技术失控，指随着技术的发展，人工智能不仅会超越人类的控制能力，还有可能摆脱人类的控制独立发展甚至反噬人类。由于当前产业大

*通信作者
修改稿收到日期：2023年7月6日

部分场景是非封闭的①，现有的弱人工智能远远无法达到控制人类的程度，暂不存在失控风险。技术的非正当应用包括无意的技术误用和有意的技术滥用2类；在封闭条件下，人工智能技术本身是中性的，是否出现误用或滥用，取决于在实际场景中具体使用的合理性（即正当与否）[1]。如果应用在正当领域，如侦查部门利用人脸识别技术抓获大量犯罪嫌疑人，产生的是正面反馈；如果应用在非正当领域，如犯罪分子利用“深度伪造”（deepfake）诈骗、勒索，则会导致隐私泄露、钱财损失等负面反馈。

在数据智能社会背景下，全球范围内治理人工智能伦理的国际硬法②缺位、国际软法③发达。同时，国际软法也面临着国际软法主体间合作不稳定、有时得不到有效实施的治理挑战。目前国内外均尚未提出针对该领域治理困境的可行建议，因而本文主要考察当前国际软法治理人工智能伦理的现状及挑战，并提出国际软法治理对策。

1 国际软法治理人工智能伦理的现状

国际软法因具有灵活高效、适用成本低，能填补硬法空白，以及方便区分治理、分层应对伦理问题的优势，其在人工智能伦理治理方面的勃兴几乎是必然的。

1.1 人工智能技术涉及的伦理问题

人工智能技术所涉及的伦理问题主要包括但不限于合理性问题、可控性问题和重大社会问题。

合理性问题，指一项人工智能技术的应用伦理上的负效应接近或超过可容忍范围[1]。数据智能社会中，人们在日常生活中使用大量智能产品；如果网络黑客通过恶意程序远程操纵无人驾驶汽车系统、人类心脏智能起搏器等，都可以轻易致人死亡[2]。在人工智能系统处理大量敏感信息和个人数据的过程中，数据被污染、泄露、滥用，不仅会影响输出结果，还可能危及人身财产安全、经济社会秩序甚至国家安全。人工智能技术的设计和生成离不开设计人员的主观意图，因此容易引发“算法歧视”和“算法偏见”。例如，美国COMPAS④量刑辅助系统加剧了种族歧视⑤。

可控性问题，指人类无法控制一种人工智能的持久存在及未来发展方向，那么它具有不可控性[1]。当前人工智能技术已具备深度学习能力，由于其存在不可解释性及不可预测性，极易产生“算法黑箱”及“欠缺透明性”等问题[3]。司法领域在确定设计、制造、使用人工智能等环节的各方主体责任及义务时有难度。例如，智能医疗助手在医疗手术过程中失误致患者手术失败、自动驾驶汽车出现决策错误导致交通事故发生，如何界定这些人工智能产品侵权主体的范围，是法律责任认定过程中尚需解决的核心问题。

重大社会问题，指人类面临的与人工智能伦理治理相关的不可逆的社会问题[1]。社会伦理方面，数据信息环境的改变使人们易陷入“信息茧房”“过滤气

① 传统的工业自动化都是封闭化，如汽车生产线；而原始的生产过程是人工完成的，是非封闭的，因为人工包含的大量因素是无法完全严格描述的，存在很多丢失变元、难解变元。将原始的生产过程改造为工业自动化过程，使其所有变元都能够完全描述并精确控制，用AI的观点看就是封闭化。封闭化的典型应用场景主要有智慧农业、物流、信息技术（IT）和制造业等。

② “硬法”（hard law）指那些能够依靠国家强制力保证实施的法规范。国际硬法主要指国际条约和国际习惯法，长期以来作为规范各国行为、确定国际权利义务关系、促进国际协调合作、应对解决各种国际问题和争端的基本依据和方法。

③ “软法”（soft law）是相对于硬法而言的，指那些不能运用国家强制力保证实施的法规范。软法这一概念最先出现在国际法领域，国际软法是软法在国际法上的延伸和发展。由于国家主权的存在，一致性的硬法规则有时难以实现，于是出现了区别于传统国际条约和国际习惯法的国际软法。

④ Correctional Offender Management Profiling for Alternative Sanctions，指替代制裁的惩教罪犯管理分析系统。

⑤ Larson J, Mattu S, Kirchner L, et al. How We Analyzed the COMPAS Recidivism Algorithm. (2016-05-23)[2023-02-07]. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

泡”和“回声室效应”⑥。它们都属于利用信息或算法将用户套牢，使其受困于狭窄的信息视野。市场经济方面，机器决策可能会产生“无用阶级”，以及平台垄断造成的阶层分化和固化，会深刻影响市场中的合作模式、竞争关系、雇佣关系、所有关系（所有权）及相关伦理规范⑦。

1.2 人工智能伦理领域软法治理的必要性

虽然人工智能技术所涉及的伦理问题很多，但其治理基本可以从2个层面来理解：① 以治理推动人工智能伦理的发展，强调运用科技治理手段、原则和工具，以治理人工智能伦理的具体问题，如人工智能技术引发的与算法公平、数据隐私和安全相关的伦理问题[4]。② 以伦理保障人工智能的治理，强调根据“理论—规则”驱动进路，将特定群体（如伦理学家和技术开发者）认可的价值取向和伦理标准程序化为伦理代码，为人工智能治理框架提供保障。如科学家或算法工程师将不同场景下（伦理学家解释的）伦理原则嵌入到算法设计中，形成对伦理观念的集体共识[4]。

作为一个综合性概念，人工智能的伦理治理最近已经形成，以治理推动人工智能伦理的发展表现为2个方面：① 世界上各国（地区）相继出台相应的法律法规来规范人工智能伦理；② 人工智能伦理相关治理机构相继建立以监督人工智能发展[4]。由各国（地区）政府及其相关部门出台人工智能技术和与人工智能技术高度相关的数据隐私方面的立法，具有法律约束力；或者由国家共同制定的国际条约和国际习惯法等，对所有的国际法主体具有法律上的强制力。此时，人工智能伦理原则嵌入上述传统法律工具之中，会展现出“刚性”的一面（硬法）。例如，2017年以来，诸多国家和地区出台人工智能治理的法律、法规、准则，立法呈现出从宏观性准则和发展战略逐渐细化至诸如自动驾驶规制、数据安全等具体层面的态势，相关人工智能伦理原则嵌入其中要求人们必须遵守⑧。

与此同时，人工智能的复杂性、不确定性、不可预见性、跨部门及跨领域多主体的特征，共同决定了对其伦理进行治理时，治理的工具必须适应发展速率更快、利益更加多元的技术不确定性环境。非常“刚性”的规则、固定模式的法律（硬法）很难跟上技术发展节奏，难以灵活规制人工智能技术所产生的社会问题。伴随人工智能技术的发展成熟，面对人工智能引发的伦理问题，此时急需在科技伦理原则和规范的指导下，发展人工智能伦理治理的新视角、新方法。

1.3 人工智能伦理领域的国际软法渊源

凡具有法律约束力的规则，如条约、习惯国际法等，均属硬法；而不具有严格的法律约束力，却能产生社会实效的规则应归类为软法，如各类指南、建议、宣言、行动守则等[5]。在国际实践中，国际硬法

⑥“信息茧房”(Information cocoons)，指在信息传播中，因公众自身的信息需求并非全方位的，公众只注意自己选择的东西和使自己愉悦的通信领域，久而久之，会将自身桎梏于像蚕茧一般的“茧房”中。“过滤气泡”(Filter bubble)，指某些大数据平台通过用户所使用的浏览历史，判断出用户偏好，向用户推荐喜爱的内容，并过滤掉异质信息，为用户打造个性化的信息世界；但同时也会筑起信息和观念的“隔离墙”，令用户身处在一个“网络泡泡”的环境中，阻碍多元化观点的交流。“回声室效应”(Echo chamber)，指在一个相对封闭的环境上，一些意见相近的声音不断重复，并以夸张或其他扭曲形式重复，令处于相对封闭环境中的大多数人认为这些扭曲的故事就是事实的全部。
解忧的百宝箱.“信息茧房”“过滤气泡”与“回音室效应”的概念、联系、区分、应对措施. (2022-04-27)[2023-06-29]. https://www.bilibili.com/read/cv16343209.

⑦ 中国发展研究基金会. 我们该如何应对人工智能带来的伦理挑战？ (2022-04-25)[2023-02-07]. https://www.cdrf.org.cn/jjhdt/5923.htm.

⑧ 全球技术地图. 国外人工智能安全相关法律法规情况. (2023-04-25) [2023-06-28]. https://baijiahao.baidu.com/s?id=1764143596130586553&wfr=spider&for=pc.

主要体现为《国际法院规约》第38条规定的国际条约、国际习惯和一般法律原则；国际软法主要体现为国际组织、多边外交会议通过的包括声明、决议、宣言、指南或者行为守则等在内的非条约式国际性法律文件[6]。与国际硬法具有强制法律效力不同，国际软法原则上不具有法律约束力但可能产生实际效果，其制订经验对国际关系及日后的国际法仍会产生一定影响[7]。譬如，《世界人权宣言》（*Universal Declaration of Human Rights*）作为多边国际法规范，虽然是软法性质的宣言，但为后来联合国通过2个硬法性质、具有法律约束力的人权公约《经济、社会与文化权利国际公约》（*International Covenant on Economic, Social and Cultural Rights*）、《公民权利与政治权利国际公约》（*International Covenant on Civil and Political Rights*）确立目标[8]。

本文在国际法领域对人工智能伦理的治理进行探讨，不涉及比较法领域的论证说明，因此本文采用国际软法一般定义——虽然缺乏国际法的法律约束力，但在国际实践中能产生实际效果的，在内容中包含有标准、规范、原则或其他行为规则的国际性法律文件。实践中人工智能伦理方面的国际软法渊源，主要来自联合国教科文组织（UNESCO）（政府间国际组织）出台的和其他国际组织（非政府间国际组织）制定的国际软法性文件等。

联合国教科文组织为推动各方遵守人工智能伦理原则，自2017年起积极制订国际软法性文件，短短4年间制订的文件在引导人工智能技术向着“负责任”的方向迈步；而其他国际组织制订的国际软法性文件，可以追溯到2005年（表1）。总体来看，关切人类福祉（“造福于人类”）、透明度和问责制是目前大多数国际软法性文件所必备的伦理要素。这些文件重点关注人的尊严，作为人工智能伦理的重要价值，关切人类福祉（“造福于人类”）的人权理念贯穿各文件始终。

其中，《人工智能伦理问题建议书》（*Recommendation on the Ethics of Artificial Intelligence*）明确表示，随着技术的发展与时俱进，对人工智能伦理的定义并不做出唯一解释；其主要原因是目前这种定义需跟随技术更迭而日臻完善，对人工智能伦理设置精确的要求有困难，只能对人工智能伦理的界定模糊不清，因此传统的“硬法”由于规则内容的精确性和固定性而暂不适合管辖。由联合国教科文组织陆续出台伦理报告、建议等“软法”以治理人工智能伦理问题，成为其他国际组织制订相关国际软法性文件的样本，文件中也都未对人工智能伦理下精确的定义。此外，西方末世论、超人文化与中国共生论、人本文化的价值观不同[2]，即使在国际法层面各方期待嵌入普世价值统一判断标准，但在构建国际硬法体系时很难完全达成有关人工智能伦理原则的国际共识。因此，国际法领域人工智能伦理的治理亟待观念革新与沟通协调，国际软法的勃兴几乎是必然的。

1.4 当前国际软法治理的优势

《人工智能伦理准则的全球格局》（*The Global Landscape of AI Ethics Guidelines*）一文中指出，“私营公司、研究机构和公共部门组织发布了人工智能伦理的原则和指导方针，围绕5项道德原则（非恶意、透明度、责任、公平正义和隐私）出现的全球趋同，这些原则具有相当大的重叠”[9]。虽然可以通过技术解决的人工智能伦理问题一般能达成全球共识，但许多技术上无法解决的伦理问题（如人工智能行业中的性别差距、可持续性问题、双重用途问题，以及劳动力流离失所等）往往被严重忽视。人工智能具有区别于其他技术的部分特点，所引发的风险已经超出监管机构对传统安全、环境和健康风险的判断，一国政府运用传统监管方法难以全面治理人工智能发展中出现的伦理问题。此外，由于人工智能本身发展速度快，涉及跨行业、政府机构、司法管辖区和利益相关者团体时，运用硬法跨主体协调监管相应的伦理问题时，因

表1 国际软法性文件中对人工智能伦理原则的具体规定

Table 1 Specific provisions on ethical principles of artificial intelligence in international soft law documents

发布机构	文件名称	发布时间	具体规定
联合国教科文组织联同世界科学知识与技术伦理委员会（COMEST）	《机器人伦理报告》（*Report of COMEST on Robotics Ethics*）	2017年12月	人的尊严、自治的价值、隐私的价值、“不伤害”原则、责任原则、行善的价值、正义的价值
联合国教科文组织第41届大会	《人工智能伦理问题建议书》（*Recommendation on the Ethics of Artificial Intelligence*）	2021年11月	相称性和不损害、安全和安保、公平和非歧视、可持续性、隐私权和数据保护、人类的监督和决定、透明度和可解释性、责任和问责
欧洲机器人研究网络（EURON）	《机器人伦理学路线图》（*The EURON Roboethics Roadmap*）	2005年	技术二重性、机器格化、人机关系的人性化、技术沉迷、数字鸿沟、技术资源获取的非平等性、技术对全球权力与财富分配的影响、技术对环境的影响
电气与电子工程师协会（IEEE）	《合伦理设计：利用人工智能和自主系统（AI/AS）最大化人类福祉的愿景》（第一版）（*Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version 1)*）	2016年12月	保护人类利益原则、问责原则、透明原则、人工智能系统或智能机器的教育与意识
生命未来研究院（FLI）	《阿西洛马人工智能原则》（*Asilomar AI Principles*）	2017年	安全性、故障透明性、司法透明性、责任、价值归属、人价值观、个人隐私、自由与隐私、分享礼仪、共同繁荣、人类控制、非颠覆性
电气与电子工程师协会（IEEE）	《合伦理设计：利用人工智能和自主系统（AI/AS）最大化人类福祉的愿景》（第二版）（*Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version 2)*）	2017年12月	人类权力、优先考虑人类幸福感、问责原则、透明原则、技术滥用意识
经济合作与发展组织（OECD）	《人工智能发展建议》（*The Recommendation on Artificial Intelligence*）	2019年5月	包容性增长、可持续发展和福祉、以人为本的价值观和公平、透明度和可解释性、稳健和安全、问责制

执法成本高、各主体观点分歧大而难有联合执法，软法此时易灵活适用以协调各方利益。

同样，与国际硬法相比，当前国际软法对人工智能伦理的治理也有优势。

（1）国际软法具有重要的优势，更为灵活高效、适用成本低。一般软法可以相对较快地被修订和采纳，而不必经历传统的政府规则制定程序。有时因不受立法机构授权的约束，几种不同的软法甚至可以同时实施，即使会产生并不相同的专业标准或者软法激增的情况，仍可以在利益相关者之间建立合作而不是对抗关系。因为在一个国际竞争激烈的时代，大部分国家政府（如美国、法国、日本）是不情愿通过先发制人的监管来阻碍新兴技术创新的[10]。各国政府会通过诸多建议、指南、行为准则、最佳实践、专业标准等国内软法，以尝试解决现实层面人工智能伦理问题⑨。国际层面更是如此，非政府组织（如国际标准化组织）由于不涉及执法机制，可以灵活高效地出台国际软法性文件，这些文件（国际软法）总体上制

⑨ 国际条约作为硬法必须遵守，这既是国际惯例，也是一国在国际社会立足的前提。硬法的订立给各个参与的主体提出了具体明确和有法律约束力的行为规范。三维度指“义务”（obligation）、“精确度”（precision）和“授权”（delegation），其中授权维度中暗含了主权的3个分支——立法权、行政权、司法权。一些国家之所以在订立硬法时望而却步，原因是硬法的授权维度要求参与成员国分割司法权。而软法可以在三维度中弱化一个或几个要素，由于这种优势而成为各成员国的选择对象。

订、实施环节的运行成本低于国际硬法。

（2）国际软法可以补充或完善硬法，通过填补正式法律空白，作为主要备用选项或治理工具[11]。国际软法不受有限机构授权的约束，因此可以解决技术引起的任何问题。而且由于没有被正式的法律机构采用，不限于特定的法律管辖区，而是可以具有国际适用性[9]。例如，国际金融研究所（IIF）发布的《数据伦理宪章》（*Data Ethics Charter*），将有助于补充（而不是取代）国家和地区立法和条例（国内层面）、国际标准（国际层面）。随着客户数据和算法的使用，《数据伦理宪章》虽然不具有强制约束力或法律地位，不建议各国政府及其金融监管部门“监管”数据伦理原则落地，但文件中IIF邀请各利益相关方将数据伦理纳入金融实践活动中，填补金融服务领域法律空白[12]。

（3）国际软法有利于人工智能伦理的区分治理、分层应对。目前，人工智能技术的发展尚处于弱人工智能阶段，大多数用途并不会产生高风险，也不需要我们面临诸如自动驾驶汽车“电车难题”的功利主义选择等类似伦理难题⑩。对于低风险人工智能应用，在公司、民间社会团体、学术专家和政府的意见下制订的软法将是促进符合道德框架和原则的创新的关键方法[13]。若软法能对人工智能伦理分级监管，对于高风险的人工智能应用，也可以成为硬法的重要补充⑪。例如，德国数据伦理委员会（Datenthikkommission）发布的《针对数据和算法的建议》（*Der Weg der gestaffelten Algorithmen*）报告，将伦理考虑分为对数据或算法系统的关注⑫，且确定了算法系统关键性的5个级别：潜在危害为零或可忽略不计的应用程序将不受监管；随着潜在危害的增加，监管负担将增加，直至完全禁止。对于具有严重潜在危害的应用程序，该报告建议持续监督。

2 国际软法治理人工智能伦理的挑战

2.1 国际软法主体间合作不稳定

鉴于人工智能应用的广泛性及其未来快速发展而导致的不确定性，截至目前非国家行为体（如IEEE、FLI、OECD等）未有国际硬法的提议与协调。当各国都选择对己有利的收益分配而不依赖具有强制效力的国际硬法来保证合作持续时，非国家行为体仅借助国际软法无法保证各国所作承诺的可信性。因此，各国间的合作是不够稳定的，这一点可根据“囚徒困境”博弈（协作型博弈）予以阐述缘由。

数字产品进行跨国数字贸易就属于“囚徒困境”博弈：各国国内施行数字贸易措施时援引世界贸易组织（WTO）或自由贸易协定（FTA）规则中的“安全例外”条款，极有可能是满足其国内数据伦理要求的；但在国际层面，这种维护各国自身利益的手段并不利于全球数字经济秩序发展。例如，美国就曾利用安全例外条款打压华为在全球市场的销售。如果各国都通过滥用“安全例外”条款禁止数字产品进入各自国内市场（即变相的贸易保护主义），将严重阻碍跨国数字贸易的发展，这是集体最差的结局；但是各国都适度开放本国的数字贸易市场，就能取得集体最优

⑩ 韩东屏．积极应对自动驾驶的“电车难题”. (2022-06-28) [2023-06-20]. https://baijiahao. baidu. com/s? id=1736840154575501899&wfr=spider&for=pc.

⑪ 例如，自动驾驶车辆对车辆(V2V)通信传递的信息能够带来潜在安全优势，可以允许近距离车辆之间的协调，以降低事故发生的可能性并提高交通流的效率。实现V2V的潜力，需要交通部门(如处理核心安全问题、避免事故)利用软法(如与车辆之间无线信息交换的程序和协议有关)进行监管。

⑫ 对于数据，该报告表明与数据相关的权利将在道德领域发挥重要作用。例如，确保个人提供知情同意以使用其个人数据解决了许多重大的道德问题。但对于算法系统，该报告表明人工智能系统可能与受影响的个人没有联系。因此，即使是没有相关权利的非个人数据也可能以不道德的方式使用，在可能造成伤害的情况下，监管是必要的。

的结果。实际对每个国家来说，最优策略仍是他国对本国开放数字贸易市场，本国则对他国实行变相的贸易保护主义。当个体理性与集体理性相悖，容易造成各国违反在 WTO 或 FTA 规则中的相关数字贸易政策承诺，即使这极可能符合各国国内的数据伦理（包括人工智能伦理）要求，因此急需国际硬法来压制各国可能采取的背弃策略。虽然过去需要具有强制效力的国际硬法介入，但在目前各国间数字经济发展的相互博弈情况下，如何发挥国际软法特有的优势，并与未来可能形成的国际硬法相辅相成、共促稳定合作，是国际软法治理人工智能伦理所面临的挑战。

2.2 国际软法有时得不到有效实施

国际软法有时得不到有效实施，反映在人工智能伦理领域，颇有争议的是涉及立法问题的问责制。人工智能的问责制，指建立具体的法律制度以说明为什么和采取何种方法能使人工智能系统的部署者和设计者担责。当人工智能系统具备了一定的行为和决策自主权后，明确责任主体就会变得愈加困难和复杂，面向各类行为主体（如设计师、制造商、运营商和最终用户等）建立的多层责任制度难以划分责任[3]。即使在跨国伦理规范范本中已经规定并描述了问责制，但是这些软法缺少正式、具体的责任划分制度，有可能在现实中得不到有效实施。

作为制订联合国全系统人工智能治理工作机构战略方针和路线图的领导机构，国际电信联盟（ITU）协调联合国各机构的参与⑬。虽与所有有关联合国实体密切合作并酌情与外部伙伴密切合作，但只能作为“促进”协调执行的机构协调中心，其执行能力有限。此外，《人工智能伦理问题建议书》中第 42 条关于“责任和问责”方面，“应建立适当的监督、影响评估、审计和尽职调查机制”；其中，“适当”的措辞本身不具有强制执行力，其实施主要靠条文本身的吸引力、各成员国对价值利益取向的共同性等加以保障。仅依靠上述推荐性、建议性的软措施，难免造成软法有时得不到有效实施的结果[7]。

3 国际软法治理人工智能伦理的对策

在国际软法发达、硬法落后的人工智能伦理领域，为了维持国际合作的稳定、促使软法内容可执行，构造国际软硬法混合治理的“中心—外围”模式、构建间接执行机制具有实践意义。

3.1 构造国际软硬法混合治理的“中心—外围”模式

3.1.1 “中心—外围”模式的提出

*在跨国法律体系中，国际硬法位居体系构造的中心，与被定义为外围的国际软法共同在国际层面构成“中心—外围”关系。*在国际软法发达、硬法落后的人工智能伦理领域，在建立各方利益表达、补偿、协调机制时，国际软法具有治理功能上的优势和事实上的效力，但这并不代表着因其具有必要性而应处于中心地位。因为国际规范和专业标准形式的软法可在不引起治外法权问题的情况下施加影响，不受颁布法律法规的政府管辖权的限制，以市场为导向激励企业制订专业标准，可使全球市场消费者来检验标准的“好坏”[14]。这种多元主体参与国际层面的治理具有协同作用，可以增强公私治理主体的协作意识，从而形成国际软硬法混合治理的“中心—外围”模式（图 1）。

*国际软法的优势领域多数情况下仅限于事务性国际关系，而不是基础性国际关系。*这是根据人工智能伦理全球治理技术性强、复杂多变的特性所决定。相较于国际硬法（传统国际法），国际软法实施起来更有弹性且方便修改。因为传统国际法调整外交、领

⑬ UN System Chief Executives Board for Coordination. A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence. (2019-06-17)[2023-02-01]. https://unsceb.org/sites/default/files/2020-09/CEB_2019_1_Add-3-EN_0.pdf.

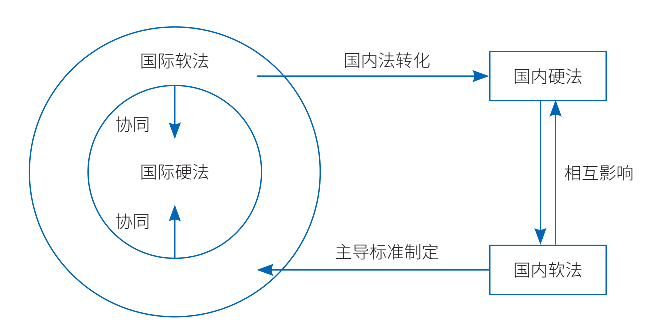


图1　国际软硬法混合治理的“中心—外围”模式

Figure 1　“Centre-periphery” model of soft-hard mixed law governance

土、承认等国家间的“共存”关系，对后来兴起的各国间有关人工智能伦理的“合作”关系，其基础性部分仍依赖稳定性强且可靠的国际硬法支撑。例如，二十国集团（G20）、OECD等重要国际组织推动各方在包含人权的“负责任的人工智能”“可信赖的人工智能”等重要概念上达成一致，其中所涉的伦理共识都必须接受执行《世界人权宣言》作为最低限度的秩序。

因此，国际软法有时体现的只是过渡形式，一旦时机成熟，仍可将国际硬法对人工智能伦理的规制提上议事日程。如人工智能应用中的自动驾驶汽车伦理，就需要根据发生的不同交通事故危险程度厘清责任归属，最后将自动驾驶汽车伦理决策判断和选择所造成的损害后果进行认定，使车辆开发者、生产者、销售者、所有者按照具备伦理原则的国际硬法的规定去遵守。因为自动驾驶汽车产品开发的时间与国际硬法规则制定的时间相似或更长，所以适用于环境变化不太快的具体技术的硬法领域[15]。

3.1.2　“中心—外围”模式的要义

在人工智能伦理的治理领域，由国际硬法和国际软法构成的“中心—外围”模式的要义包括4个方面。

（1）国际软法向国内硬法的渗透，尤其是技术标准经由国内法转化而具有约束力。例如，国际标准化组织（ISO）和国际电工委员会（IEC）制订的《隐私信息管理要求与指南》（*Privacy Information Managment—Requirements and Guildlines*，ISO/IEC27701）作为技术软法，与欧盟《通用数据保护条例》（*General Data Protection Regulation*）第4章第5节“行为为准则和认证部分”标准相衔接，经由欧盟成文法《通用数据保护条例》转化并推动标准互认。成文法规定嵌入数据保护规范的条件和具体方法，能进一步明确认定技术标准的法律效力[16]。

（2）通常有意向发展人工智能技术的政府，会鼓励企业、用户群体、非政府组织、科研机构等多元监管主体积极参与相关行业协会对人工智能伦理的研讨，促进国内软硬法相互影响、相互转化。例如，欧盟陆续出台指导性文件《可信AI伦理指南》（*Ethics Guidelines for Trustworthy AI*）、《算法责任与透明治理框架》（*A Governance Framework for Algorithmic Accountability and Transparency*），倾向于规定人工智能伦理的混合性条款，即借鉴软法的开放协调程序以定期修订条款，通过模糊硬法与软法的边界，使软法和硬法的双重手段在实践中优势互补、共同推动人工智能伦理建设进程。

（3）各国国内企业、学术界和政府间国际组织等，可以出台各类围绕人工智能的特定用途的文件，以期将各国国内技术治理偏好向国际软法渗透，争取国际标准制订的话语权。专业标准可能需要数年时间来制订，因此并不总是比硬法更具有速度优势。然而，与硬法相比，专业标准通常是专门为促进行业内的创新而制订的，并且可以随着技术的发展而相对快速地更新。例如，IEEE正在制订主题包括“自治系统的透明度”“解决透明度、问责制和算法偏差的认证方法”的人工智能标准，以及人工智能对人类福祉的影响[17]；或者像ISO正在与IEC合作开发一连串以人工智能为要点的标准。而一旦创建了这些专业标准，数字产业的多方利益相关者或现有私营机构就可以在全球范围内传播和执行，逐渐影响乃至主导国际标准

制订。

（4）即使国际软法的影响力与国际硬法势均力敌，也不会从大局上影响国际软硬法混合治理的“中心—外围”模式。国际硬法具有法律约束力，通常在实践中的影响力往往会比只有事实效力的国际软法更大[18]。未来由各国共同制定的国际硬法，相比由其他非国家行为体制订的国际软法，其国际合作更加稳定且正当性基础要更为坚实。例如，容易产生高风险的自动驾驶汽车技术，由于关切人类福祉（如生命健康权等），即使国际软法具有立法成本低等优势，还是需要国际硬法确定其基础伦理规范，实施阶段硬法成本反而容易比软法低。

综上，在企业及行业协会拟议人工智能伦理准则时，可以参考技术标准在相同议题上的规范，尤其是透明度、问责制、关切人类福祉等普遍性技术议题。例如，人工智能伦理中透明度原则的设立，可以采取国际软、硬法混合治理的“中心—外围”模式：利用硬法（由国家共同制定的法律）规制来解决更高层次、普适性较强的数据/算法/算力/知识突出问题，进行较强力度的底线约束，可以提高人工智能的可靠性；同时，利用软法（非国家行为体制订的规则）解决具体应用场景的个性化问题，允许比立法机构更接近技术的专家开发软法框架，提供有一定容错空间的韧性规范以指导技术的发展。

3.2 构建间接执行机制

3.2.1 建立国际治理协调委员会

Marchant和WallachGary[19]从国际角度提出了一个协调实体，可称之为“治理协调委员会”（governance coordinating committee）。若在人工智能伦理领域成立一个国际治理协调委员会，该委员会“位于政府之外，但包括政府代表、行业、非政府组织、智库和其他利益相关者的参与”，且“该实体不会寻求复制或取代许多致力于开发人工智能治理方法的组织，而是提供一种协调功能……确保所有不同的参与者相互联系，了解并响应彼此的提案，同时识别现有计划中的差距和不一致之处”[9]。因此，多元主体参与国际治理协调委员会，有利于其在定义伦理准则时具有可信度，能促成软法“硬化”以增强软法约束力和执行可能性[20]。

人工智能机构间工作组（IAWG-AI）由联合国教科文组织和国际电信联盟共同领导，通过支持行政首长协调会和方案问题高级别委员会关于人工智能伦理的工作，寻求在人工智能相关问题上与相关机构间和多利益相关方机制合作⑭。该工作组重点关注伦理领域，可在其基础上设立国际治理协调委员会，这既可以降低监督机构的设立成本，也可以提高国际影响力。各种类型的软法（包括伦理标准、指南、行为准则和原则等）在规范人工智能伦理时，国际治理协调委员会可以依赖惩罚手段、社会舆论监督和技术专家的技术评价等予以保障实施，间接影响和引导相关责任主体承担应有的责任（图2）。

成立国际治理协调委员会，不仅可以拥有在人工智能机构间工作组中保持“技术中立”的专家资源，还可以提高行政首长协调会和方案问题高级别委员会所管辖事务的治理效率。国际治理协调委员会作为非国家行为体，制订的国际软法虽然无法像国际硬法那样得到国家强制力单独的（主要依靠主权国家自身的力量）或集体的（在某些情况下依靠国际社会的集体力量）支持，但可辅之以非正式和非集中化的间接执行机制，取得相应的实效[18]。

3.2.2 非正式和非集中化的间接执行机制

在跨国法律体系框架下，国际软法作为治理工具

⑭ UN System Chief Executives Board for Coordination. Inter-Agency Working Group on Artificial Intelligence. (2021-03-01)[2023-02-01]. https://unsceb.org/inter-agency-working-group-artificial-intelligence.

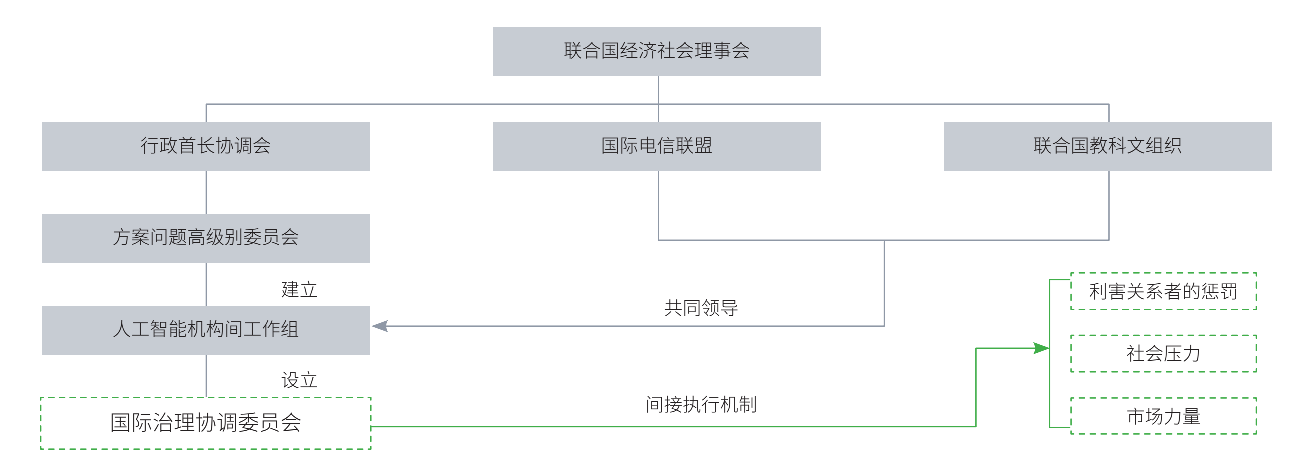


图 2　联合国人工智能治理工作机制及构建间接执行机制设想

Figure 2　United Nations working mechanism for AI governance and assumption of constructing indirect implementation mechanisms

绿色虚框内容在实践中并未成形,是构建间接执行机制的设想

The content of green virtual boxes has not been formed in practice, and it is a tentative idea to construct indirect implementation mechanisms

不具有法律约束力和强制执行效力，但具有实施的效果。许多非国家行为体制订的国际软法虽然没有强制执行力保证实施，无法像传统国际法对违法者诉诸国际司法或仲裁机构和进行报复等，但决不能否定国际软法具有自我管制和自我约束的功能，外部力量的作用仍能使其得到事实上的遵行。依据博弈论，通常情况下依靠各软法主体的“自我实施”，不需要有第三方强制性力量的保障和集中化的权力机构，辅之以非正式和非集中化的执行机制，如利害关系者的惩罚、社会压力及市场力量，就能促使国际软法得到事实上的效力[18]。

（1）利害关系者的惩罚提高违法成本。一般基于硬法工具的威慑，利害关系者才会遵从软法。但非国家行为体对不遵守软法的利害关系者进行处罚，提高其违反软法的潜在成本，也可以形成事前的积极威慑并达到预防的目的。例如，一国违反世界银行制订的《外国直接投资待遇指南》，不但无法从世界银行获得担保和贷款，反而还会被此权威性的国际经济组织宣布为投资环境欠佳的国家，严重影响该国未来吸收外资的能力[18]。建议：国际治理协调委员会设置提高违反软法成本的惩罚措施，间接提高国际软法的执行力。

（2）社会压力促进声誉机制。违反国际软法的主体极易受到来自国际社会的压力，国家、跨国非政府组织、政府间组织等多利益攸关方的谴责都会使违法者名誉扫地、颜面尽失。有些国际软法是名义上的软法、实质上的硬法，极大地损伤违法者的经济或商业利益，其惩戒制裁效果不亚于对其不利的国际仲裁裁决或司法判决。这种“群起的羞辱”间接促使各国遵守国际软法、维护其在国际社会上的信誉，避免因不执行国际软法而失去国际合作伙伴信赖的不利影响。建议：国际治理协调委员会与联合国教科文组织、G20、国际电信联盟等国际组织积极讨论、推动多边共同监管的行动倡议或伦理规则，探索人工智能伦理的共识与差异[14]。声誉机制围绕非国家行为体的倡议或规则执行，要求对破坏倡议或规则的国家进行惩罚，对维护倡议或规则的国家进行褒奖，从而使得各国为了维持国际社会良好声誉，迫于国际社会压力而承担国际社会责任。

（3）市场力量催生“网络效应”。当接受统一准则

或标准的国家聚合，且使用该准则或标准的国家增加时，准则或标准对它们的价值也相应增加，从而吸引更多的国家自发聚合形成“雪球效应”，不断扩大遵守软法的群体规模，增强软法约束力。不接受的国家就会被边缘化，接受的国家则会获益越来越多，由此推动各国接受这些国际软法（统一准则或标准）。例如，ISO 制订了纳米技术风险管理的国际标准，作为安全处理纳米材料的准监管标准。虽然这些标准不能直接执行，但许多国际合同和一些保险公司要求遵守适用的ISO标准（或其等效标准），从而提供间接执行机制[21]。又如，国际治理协调委员会学习国际干细胞研究学会制订的准则，为限制某些类型（反伦理）的研究而制订研究指南，并为其他类型的研究提供伦理保障。虽然不能直接执行，但这些指南为干细胞研究人员设定了专业期望，并且可以由研究机构、资助机构和要求科学家遵守的科学期刊间接执行[21]。**建议：**① 建立起多元化的参与路径，推动企业、行业协会、专家学者等多层次国际对话，利用科学期刊分享最佳实践以支撑标准化工作；② 不同国家人工智能伦理规则在不同文化背景下的协调互动、交流碰撞[14]，在国际交往中逐渐形成不成文的准则或标准（即国际惯例），从而促使各国在国际实践中去遵守约定俗成的人工智能伦理规范。

参考文献

1 陈小平. 人工智能伦理导引. 合肥: 中国科学技术大学出版社, 2021.
Chen X P. An Interdisciplinary Guide to Artificial Intelligence Ethics. Hefei: University of Science and Technology of China Press, 2021. (in Chinese)

2 高奇琦. 人工智能治理与区块链革命. 上海: 上海人民出版社, 2020.
Gao Q Q. Artificial Intelligence Governance and Blockchain Revolution. Shanghai: Shanghai People's Publishing House, 2020. (in Chinese)

3 莫宏伟, 徐立芳. 人工智能伦理导论. 西安: 西安电子科技大学出版社, 2022.
Mo H W, Xu L F. Introduction to Ethics of Artificial Intelligence. Xi'an: Xi'an University of Electronic Science and Technology Press, 2022. (in Chinese)

4 Xue L, Pang Z J. Ethical governance of artificial intelligence: An integrated analytical framework. Journal of Digital Economy, 2022, 1(1): 44-52.

5 胡玉婷. 论软法与硬法在多维界分中的渐变——以《巴塞尔协议》为视角. 东方法学, 2014, 38(2): 106-114.
Hu Y T. On the gradual change of soft law and hard law in multi-dimensional boundary—From the perspective of *Basel Accord*. Oriental Law, 2014, 38(2): 106-114. (in Chinese)

6 沈伟, 冯硕. 全球主义抑或本地主义: 全球数据治理规则的分歧、博弈与协调. 苏州大学学报(法学版), 2022, 9(3): 34-47.
Shen W, Feng S. Globalism or territorialism: difference, game and coordination of global data governance rules. Journal of Soochow University (Law Edition), 2022, 9(3): 34-47. (in Chinese)

7 罗豪才. 软法与公共治理. 北京: 北京大学出版社, 2006.
Luo H C. Soft Law and Public Governance. Beijing: Peking University Press, 2006. (in Chinese)

8 Gruchalla-Wesierski T. A framework for understanding “Soft Law”. McGill Law Journal, 1984, 30(1): 37-88.

9 Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 2019, 1(9): 389-399.

10 Marchant G. “Soft law” governance of artificial intelligence. (2019-01-25) [2022-08-30]. https://escholarship. org/content/qt0jq252ks/qt0jq252ks_noSplash_1ff6445b4d4efd438fd6e06cc2df4775.pdf?t=po1uh8.

11 Brookings. How soft law is used in AI governance. (2021-05-27)[2022-08-30]. http://www.brookings.edu/techstream/how-soft-law-is-used-in-ai-governance/.

12 Bailey N, Ferenzy D, Carr B. Data ethics charter. (2021-06-07)[2022-08-30]. https://www.iif.com/portals/0/Files/content/Innovation/06_07_2021_iif_data_ethics_charter.pdf.

13 Eliazàt A. Soft law as a complement to AI regulation. (2021-07-07) [2022-08-30]. https://adolfoeliazat. com/2021/07/07/

soft-law-as-a-complement-to-ai-regulation/.

14 罗豪才. 软法的理论与实践. 北京: 北京大学出版社, 2010.
Luo H C, et al. The Theory and Practice of Soft Law. Beijing: Peking University Press, 2010. (in Chinese)

15 程莹, 姬祥. 如何在人工智能治理中使用软法? . (2021-07-13) [2023-02-05]. https://mp. weixin. qq. com/s/AP8_ptSy_7Hs_unstZFKrA.
Cheng Y, Ji X. How to use soft methods in artificial intelligence governance?. (2021-07-13) [2023-02-05]. https://mp.weixin.qq.com/s/AP8_ptSy_7Hs_unstZFKrA.(in Chinese)

16 鲍坤. 数据平台下个人数据保护规则形态的优化——从软法对硬法的嵌入谈起. 中国科技论坛, 2022, (3): 156-165.
Bao K. Optimize the form of personal data protection rules under data platforms—From soft law embedded in hard law. Forum on Science and Technology in China, 2022, (3): 156-165. (in Chinese)

17 Cihon P. Standards for AI governance: international standards to enable global coordination in AI research and development. (2019-04-01)[2023-02-06]. https://www.fhi.ox.ac. uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

18 徐崇利. 全球治理与跨国法律体系:硬法与软法的“中心—外围”之构造. 国外理论动态, 2013, (8): 19-28.
Xu C L. Global governance and transnational legal system: The “Central-Peripheral” structure of hard law and soft law. Foreign Theoretical Trends, 2013, (8): 19-28. (in Chinese)

19 Marchant G, Wallach W. Coordinating technology governance. Issues in Science and Technology, 2015, 31(4): 43-50.

20 莫志. 上市公司环境、社会和治理信息披露的软法实现与强化路径. 江西财经大学学报, 2022, (2): 116-129.
Mo Z. The implementation and strengthening path of soft laws for ESG information disclosure of listed companies. Journal of Jiangxi University of Finance and Economics, 2022, (2): 116-129. (in Chinese)

21 Marchant G E. ‘Soft Law’ mechanisms for nanotechnology: liability and insurance drivers. Journal of Risk Research, 2014, 17(6): 709-719.

International soft law governance of artificial intelligence ethics: Current situation, challenges and countermeasures

ZHU Mingting* XU Chongli

（School of Law, Xiamen University, Xiamen 361005, China）

Abstract Artificial intelligence (AI) technology not only rapidly empowers economic and social development, but may also trigger many ethical issues highly related to the characteristics and development of AI technology itself. The rise of international soft law in the field of AI ethical governance is almost inevitable due to its flexibility, efficiency, low application cost, ability to fill the gap in hard law, and convenience in distinguishing governance and layered response to ethical issues. Under the current situation of developed international soft law and outdated hard law in this field, faced with the governance challenge of unstable cooperation among subjects of international soft law and sometimes unable to be effectively implemented, the governance model has gradually changed towards a combination of soft law and hard law, and the "hardening" of soft law, in order to improve the binding force and enforcement possibility of soft law. It is suggested to construct a "centre-periphery" model of mixed international soft and hard law governance, and construct an indirect enforcement mechanism to improve the international soft law governance strategies of artificial intelligence ethics.

Keywords artificial intelligence ethics, international soft law, mixed governance, indirect enforcement mechanism

朱明婷 厦门大学法学院国际法专业2020级博士研究生，厦门大学法学院网络空间国际法研究中心研究助理。主要研究领域：网络空间国际法、数据伦理、金融科技等。E-mail: 12920200155834@stu.xmu.edu.cn

ZHU Mingting Ph.D. candidate in international law at School of Law, Xiamen University, and Research Assistant at the Center for International Law in Cyberspace at School of Law, Xiamen University. Her main research covers international law in cyberspace, data ethics, financial technology, etc. E-mail: 12920200155834@stu.xmu.edu.cn

徐崇利 厦门大学法学院教授、博士生导师，厦门大学国际法学国家重点学科学术带头人之一。主要研究领域：国际经济法、国际私法及国际法基本理论研究。E-mail: clxu@xmu.edu.cn

XU Chongli Professor and Doctoral Advisor of School of Law, Xiamen University, and one of the academic leaders of National Key Disciplines of International Law of Xiamen University. His main research covers international economic law, private international law, and basic theoretical research in international law. E-mail: clxu@xmu.edu.cn

■责任编辑：岳凌生

*Corresponding author